Relationship between a set of variables and miles per gallon (MPG)

Jagdeep S. Sihota

Executive Summary

The objective of the study was to look at a data set of a collection of cars and exploring the relationship between a set of variables and miles per gallon (MPG). We were particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG ?
- Quantify the MPG difference between automatic and manual transmissions

While initial analysis shows that manual transmissions were better for MPG, but regression analysis showed MPG mostily depends on weight and horsepower and transmission was not major factor in determining MPG.

Exploratory Analysis

Data The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). A data frame with 32 observations on 11 variables (See Table 1.1 in the Appendix for more details).

Analysis Figure 1.1 shows relationship amoung all the variables, but narrow down data box plot is created figure 1.2 The box plot clearly showed that there was a difference in MPG depending on transmission type. Automobiles with a manual transmission had a higher MPG than those with an automatic.

A correlation test of association was conducted to explore this relatinship further, see Table 1.2. The results showed that MPG was most highly correlated with weight (-0.87), the number of cylinders (-0.85), displacement (-0.85), and horsepower (-0.78), although these were all negative correlations, meaning they had a negative impact on MPG levels. The results also showed that MPG was positively correlated with transmission (0.60), meaning a manual transmission was correlated with higher levels of MPG

Statistical Modeling

To address the main research question of the relationship between MPG and transmission type, we used simple least squares regression, with MPG as the outcome variable (Y) and transmission as the predictor (X). Table 2.1 shows the results of this regression analysis. The coefficient estimate for manual transmission is 7.24, which tells us that automobiles with a manual transmission get about 7.24 more gallons per mile than those with an automatic transmission. This finding is statistically significant as $p < 0.05$ and the 95% confidence intervals for the estimate were (3.64, 10.85). The multiple R-squared value is only 0.35, which indicates that only 35% of the variance in MPG is explained by the type of transmission, leaving 65% of the variance unexplained.

A better model of MPG would include additional correlated variables to reduce the unexplained variance. Table 2.2 shows the results of Model 2, which includes additional variables that were highly correlated with MPG.Model 2 significantly improves the R-squared value to 0.86, but if we remember from our correlation analysis, some of these variables were correlated to each other. To correct for multicollinearity and high VIF, we removed the redundant variables that were highly correlated with other predictors (cyl and disp). After removing these variables, we ran a third regression model with the remaining variables, see Table 2.3.

In Model 3, the R-squared value remains the same (0.85), even after removing two variables that were highly correlated to MPG. This model, however, contains predictor variables that are not statistically significant ($p > 0.05$), including transmission type (amManual). This tells us that when holding the other variables constant, transmission type is not a factor in determining MPG. The reason this occurs is because transmission is highly correlated with the weight variable (-0.69), meaning that cars with manual transmissions are lighter than those with automatic transmissions, see Figure 3. This relationship with weight is why transmission was a statistically significant predictor of MPG in Model 1, but in reality this is only a byproduct of it's correlation with weight. Hence, the best model for MPG should not include transmission as a predictor, and only include variables which are statistically significant ($p < 0.05$), see Table 7.

Finally, in Model 4 we are left with two statistically significant predictor variables, weight and horsepower. The coefficient for weight is -3.88, meaning that on average and with horsepower being held constant, a one unit increase in weight (1000 lbs) equates to a -3.88 lower MPG. This coefficient is statistically significant with a $p<0.05$ and the 95% confidence intervals are (-5.17, -2.58). The coefficient for horsepower is also negative at -0.03, which indicates that on average and holding weight constant, a one unit increase in gross horsepower will result in a .03 loss in MPG. This estimate is statistically significant ($p<0.05$) with 95% confidence intervals of (-0.05, -0.01). The R-squared value of this model is 0.83, which means that 83% of the variance in MPG is explained by the weight and horespower of the automobile.

Diagnositc Plots

To verify the conditions of validity of our final model we can look at some diagnostic plots of the residuals. The plot of Residuals vs. Fitted values is slightly U-shaped which shows that there may be some non-linearity in the relationship between MPG and one of the variables in the model, suggesting that adding a quadratic term may improve the fit of the data, see Figure 4.

After adding quadratic terms for weight and horsepower, we see that the new coefficients are both statistically significant ($p < 0.05$) and the R-squared value (0.89) was the highest among all models, see Table 8. Furthermore, if we look at the diagnostic plot for Residuals vs. Fitted values for this new model, we see that their is no longer a discernible pattern in the variance of the residuals in relation to the fitted values, see Figure 5.

Conclusions

Appendix

library(xtable) name <- names(mtcars) description <- c("Miles/(US) gallon", "Number of cylinders", "Displacement (cu.in.", "Gross horsepower", "Rear axle ratio", "Weight (lb/1000)", "1/4 mile time", "V/S", "Transmission (0 = automatic, 1 = manual)", "Number of forward gears", "Number of carburetors") dataset <- cbind(name, description) print(xtable(dataset, caption = "Motor Trend Data, 1974"), comment = FALSE, type = 'latex')